COMP-598 Final Project

Xuanzi Wang, Sierra Schena, Jaylene Zhang

McGill University
xuanzi.wang@mail.mcgill.ca, sierra.schena@mail.mcgill.ca, yihan.zhang2@mail.mcgill.ca

Introduction

COVID-19 has impacted the lives of millions around the world, infecting hundreds of millions worldwide as well as producing over 5 million deaths [7]. Scientists have banded together across different nations to produce a vaccine in hopes of eradicating the virus, with several being approved and distributed as soon as one year after the start of the pandemic. Despite this, many continue to be skeptical of receiving the vaccine, and the pandemic has continued to spread over another year.

In this project, our team was tasked with understanding the discourse surrounding COVID-19 in Canadian social media, and in particular, hesitancy surrounding vaccines. Our main goals with this project were to uncover: the main topics discussed regarding COVID and how people are engaging with these topics with respect to if the response is generally positive, negative or neutral with respect to these topics. The social media platform used for this extraction is Twitter, however due to limitations with Twitter's API, instead of limiting the geolocation to Canada only, we simply extracted any tweets in English and will infer that these results are similar to what we would find if we were to look only at Canada.

The main topics that we discovered were as follows: 'vaccine', 'virus', 'government', 'daily life', 'society' as well as an 'other' category for the tweets for which there was no main topic of interest. A further description of the topics can be found in the Results section. We have found that among these topics, people are most keen to talk about the topics of vaccine and virus when commenting on COVID-19, as well, most of the tweets have either a neutral or negative sentiment. Moreover, we have generated the words that are most related to specific topics by computing the tf-idf scores in order to have a more clear insight into each topic. Our results indicate that people have talked about vaccines the most and it is the topic that possesses the highest proportion of positive sentiment. However, the general sentiment over the distribution is topics is overwhelmingly negative.

Data

For this task, we originally collected about 1400 tweets over the span of 3 days - November 30, December 1 and December 2, 2021. We were required to compile 1000 tweets in total, however collected extra to ensure that no tweets were duplicated and that all of the tweets are closely related to COVID-19. These tweets were collected using the tweepy library in Python. Our query collected tweets that were only in English, not retweets and that contained any of the following words (irrespective of case): 'covid', 'vaccination', 'moderna', 'pfizer' or 'astrazeneca'. After the tweets were collected, we removed all duplicates and kept 1000 relevant tweets to then perform our open coding and subsequent annotations.

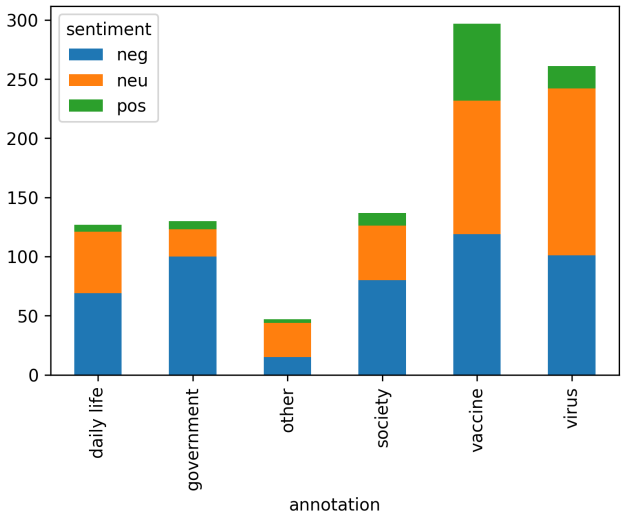

Figure 1: Counts of each topic segmented by sentiment.

As can be seen in Figure 1, the most frequently used categories were 'vaccine' and 'virus' amassing almost 300 tweets per topic. The categories of 'daily life', 'society' and'government' accounted for approximately 100 tweets each. Lastly, the smallest category with about only 50 tweets is the 'other' category. We have also tried to avoid categorizing tweets as 'other' since it would be less informative, as can be seen by its total count in relation to the other categories. As well, we wanted to ensure that the rest of the more informative categories were approximately equally

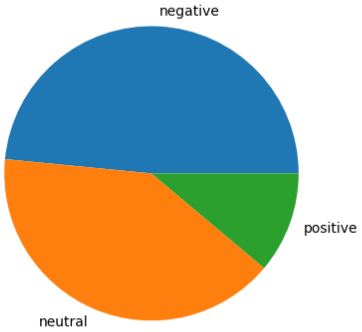

Figure 2: Proportion of each sentiment across all topics.

represented so that our tf-idf analysis would not be biased by lack of data in certain categories. If we instead look at the distribution of sentiments over these topics, we observe that tweets are overall rarely positive, with negative or neutral sentiments having the majority in each topic group. Interestingly, the most evenly distributed topic with respect to sentiment is the 'vaccine' topic. As well, we observe that from Figure 2 that the general sentiment felt towards these topics relating to COVID-19 is negative.

Methods

- Data collection: We collected 1400 tweets from Twitter using the tweepy library. Firstly, we created a developer account in order to work with the Twitter API. We followed the following instructions from [5] to collect an equal amount of tweets over a period of 3 days. We defined the query so that it filters the tweets that contains certain COVID-related key words as mentioned in the Data section, and are not retweets. We have also extracted the language field of each tweet to make sure all the tweets are written in English.
- Topic design: We conducted open coding on the first 200 tweets we collected in order to design appropriate topics. We firstly determined more detailed topics for each tweet, such as the number of new cases or criticism of the vaccine. Afterwards, we attempted to make the categories more broad so that they could include several detailed topics that we previously summarized. For example, the topic 'vaccine' could include detailed topics: criticism or advocacy of the vaccine, as well as people's personal experiences after receiving their vaccine. After merging topics together, we then limited the total number of topics chosen to be those with the most occurrences. The topics that rarely appeared were removed and those tweets were classifies as 'other', alongside any other tweet that did not clearly relate to one of the main topic groups. In this way, we ended up with a total of 6 topics: 'daily life', 'vaccine', 'virus', 'society', 'government', and 'other'.
- Data annotation: We annotated all of the tweets manually with one column to denote the topic associated to the tweet and another for the overall sentiment of the tweet. To decrease subjectivity, we made strict definitions for the topics. We also ensured to collect enough tweets so that only a small number of tweets would not fit into our defined topic groups, so that the 'other' category would be rarely used.
- tf-idf: Before computing the tf-idf scores for each topic, we cleaned and tokenized the tweets. We removed emojis in the text and replaced punctuation characters with a space, followed by splitting all the text into alphanumeric word tokens and removing stopwords. Once this was completed, we iterated through each of the six topic groups and counted how many times each word showed up in a specific topic. In addition, we also computed the number of times a word was mentioned among all six topics. We then calculated the tf-idf scores by multiplying tf(w, topic) and idf(w) where $idf(w) = log_{10}(6/w)$ since we have 6 topics. Finally, we obtained the top ten words for each category after sorting the tf-idf scores in descending order.

Results

The 6 topics selected were determined through an open coding of a subset of 200 tweets. We tried to choose topics that would not have much overlap in interpretation, as well as cover a wide range of issues that would each cover a sizeable portion of the tweets. The definition and characterization for each of the chosen topics is as follows:

- daily life: This topic encapsulates tweets where the author describes the impact of COVID-19 on their daily lives. For example, many tweets were talking about the travel restrictions, mask mandates or other consequences to their routines or schedules. Especially, at the time that the tweets were extracted, it was becoming close to the holiday season and many people were tweeting about not being able to see family, derailed travel plans or planning Christmas parties.
- vaccine: This topic includes people's attitudes towards the COVID-19 vaccine and their vaccination experience. Some positive tweets include a call for everyone to get vaccinated, as well as people's positive experiences with getting their doses, whereas some negative tweets surround the potential risks of vaccination or its true efficacy. The newly developed variant, Omicron, has also generated some buzz surrounding vaccination: whether or not to have a booster shot.
- virus: This topic involves the discussion of the spread of COVID-19 itself as well as its variants, such as Omicron, and any relevant news such as current cases in a certain district. People have focused greatly on the new virus, tweeting about the new risk it imposes in terms of rendering vaccines ineffective and the number of new cases associated with it, etc. There were also a lot of tweets

comparing COVID-19 with other diseases such as the flu which were included in this category.

- society: This topic pertains to how the pandemic has impacted society. Tweets include subjects such as social advocacy and the perception of the general public with respect to opinions surrounding COVID-19. Other examples include the financial impacts caused by the virus, where some tweets have mentioned people having lost their jobs and how the virus has slowed the economic growth.
- government: This topic covers people's perspectives towards governments, policymakers and their initiatives in reaction to COVID-19. People hope that the government could effectively control the pandemic, thus people tend to criticize the government if the virus is out of control. There are many tweets that we collected where people were criticizing the government for either a lack of policies in place to protect the citizens, for having too many restrictions, as well as comparing the practices of different governments.
- other: This topic includes all of the other tweets that could not be classified into the previous categories. Even though these tweets could not be categorized into previous topics, they are closely related to COVID-19. We have collected more tweets than needed in order to reduce the size of the 'other' category, so that most of the tweets are informative.

In terms of the engagement of each topic, we see from Figure 1 that most people are keen to talk about the vaccine and the virus when discussing COVID-19. We observe that the number of tweets that focus on daily life, government or society are lower.

In addition, we have classified each tweet to one of the following sentiments: negative, positive or neutral. The sentiment for each tweet represents the attitude of the author towards the topic. We observe that most people hold negative attitudes across all topics from Figure 2 as well as for each topic individually as can be seen from Figure 1. The number of people who hold neutral attitude is slightly lower, while the proportion of tweets that are positive is the smallest.

Moreover, we computed tf-idf scores for the keywords present in each category. Our goal is to find the most important words in each category and have an insight into people's main concerns in different areas. The results are as follows:

- **daily life**: return, departure, holiday, past, days, results, pass, plan, hope, party
- **vaccine**: pfizer, doses, vaccine, effects, boosters, risk, booster, shots, heart, study
- **virus**: confirmed, variant, risk, jump, denmark, omicron, study, cdc, according, positive
- **society**: pm, happening, fight, market, ignoring, healthcare, risk, support, countries, misinformation
- **government**: voters, gop, masked, donald, lame, duck, presidency, catastrophe, ignorance, diplomacy
- **other**: war, rogan, reported, major, available, party, lied, provides, caught, negative

Discussion

Topic engagement represents how much people care about each topic. The proportion of negative, positive or neutral sentiment shows people's attitudes towards these topics, and the tf-idf results give us an insight into the specific keywords that are commonly used in reference to these topics. For the most representative keywords, we will discuss potential reasons as to why it is frequently mentioned by people with evidence from what has happened in Canada over the span of the pandemic to try and explain its significance in its respective topic group.

Society

With respect to the society topic, 'healthcare' in the top word list gives us an insight that many people have opinions about how the medical system has changed since the beginning of the pandemic. As seen from Figure 1, we also observe that the response in the society category is mainly negative, thus we can observe that people are mostly unhappy with these changes.

In Canada, there has been a large shift to virtual care during the pandemic, enabling patients and physicians to connect safely at a distance to reduce the spread of COVID-19. [1] However, although the online format guarantees safety and promptness, it is still regarded as an inappropriate approach to some specific health concerns by a considerable amount of Canadians. It is believed that the overall quality of in-person visits is better for the majority of Canadians. In general, fewer people sought care for common concerns, such as abdominal pain or colds and flu, or even for more significant concerns as well. Recently, concerns rising over Omicron's impact in healthcare system may increase the stress on health organizations again. Major concerns include the shortage of beds, specialized equipment, medical professionals available in the hospitals and challenges on tertiary care. [3]

Other keywords which also allude to the negative sentiment that is majorly found within this topic are 'fight', 'ignoring' and 'risk', which suggest some instability between people during this time. The keyword 'support' combined with the overall negative sentiment of this topic group may allude to the lack of support that some may be feeling. 'Market' being found in the list of top words may be linked to people's concerns over how the virus has impacted financial markets and the economy. Lastly, the word 'misinformation' being found in this category is quite telling of how people feel about how society is reacting to the pandemic, feeling as though a lot of information is being misconstrued and manipulated.

Vaccine

This topic group was the most controversial, as this topic group is the most equally distributed between sentiments. We see from the list of top words associated with vaccine

that ‘pfizer’ is the most important word in this topic. Pfizer is one of the most widely accepted COVID-19 vaccines, and from the analysis, we could conclude that among all kinds of vaccines, Pfizer is the one that people talk about the most, both positively and negatively.

In this topic, ‘booster’ is also an important word. During annotation, we found that most people were curious about whether or not to get a booster shot. We could conclude from this word that as new variants come out, more people are focusing on whether or not to get a booster dose.

The effects and risks associated with vaccination have been some of the biggest factors that influenced people’s decision to get vaccinated, as can be seen from the top words list. During annotation, we found that among all of the tweets that were focused on vaccines, most of the negative ones spoke about the risks involved. In particular, we see ‘heart’ in the top words list. Indeed, there is evidence showing that the vaccine may cause damage to the cardiovascular system. According to the Government of Canada[2], side effects have occurred for 28,825 people and 6,581 cases were serious in Canada. Especially, more than 81% of the tweets that mentioned ‘heart’ have negative sentiment. We could infer that the heart risk has stopped many people from getting vaccinated against COVID-19.

We see from Figure 1 that out of all of the topics, most of the positive tweets were talking about vaccines, which means that many people consider vaccine as a positive approach in fighting the virus. In particular, more than 17% of the tweets are positive within the vaccine topic. However, approximately one third of the tweets are negative demonstrating that this is not how the majority feel, at least in the sample that we have collected.

Virus

More than 1/4 of the tweets are related to the virus itself as well as its variants. In particular, this topic possesses the largest proportion of neutral sentiment since many of the tweets corresponding to this topic are news-based, i.e. tweets reporting how the virus is evolving. During the annotation process, we observed that the top words ‘confirmed’, ‘jump’, and ‘cdc’ were frequently mentioned in the news tweets reporting about the newly confirmed cases in different districts and age groups.

It is noticeable that ‘variant’ and ‘omicron’ are also important subjects in this topic group. People have great concerns about the new variants since they may cause more severe illness or reduce vaccine effectiveness. There have been news stating that in Canada, Omicron infections are on the rise. We also know from the vaccine topic that people are considering the booster shot as the new variants are spreading rampantly and the protection of the vaccine in two doses decreases. Concerns regarding these issues can explain the large negative sentiment found in this category where approximately 38.7% of the tweets are negative and only 7% are positive. We observe that this is greatly different from the proportions of sentiments found in the vaccine topic. We therefore infer that compared to vaccines, less people hold positive attitudes towards this category since at least some people regard vaccines as a solution to limiting the spread of the virus, whereas the development of the virus into new variants is negative.

Daily Life

The pandemic has greatly affected people’s lives and many people have been looking forward to returning back to their normal lives. Indeed, from the list of tops words we see words like ‘return’, ‘past’, ‘hope’ as well as ’departure’, which suggests how travelling plans were disrupted during the pandemic. Also, people are complaining about the mandatory quarantine, the varying border restrictions, and the covid-19 test required before crossing the border, which strictly limits international travels. Moreover, as these tweets were collected at the beginning of December, many tweets mentioned ‘holiday’ and ‘party’ as people are hoping to be able to celebrate during this time. Tweets with a hopeful sentiment about being able to celebrate and get together with friends and family were however the minority as can be seen in Figure 1. It seems as though the majority of people are showing negative sentiment, as can be partly explained through the hesitancy they feel around large gatherings at this time. Many are worried that the number of cases may rise during the holiday season. From previous statistics as can be seen from Figure 3, large spikes in cases of COVID-19 have been found during and immediately after the holidays. Specifically, the largest increase in the number of cases occurred around the time period between last Christmas and the new year. Thus, a lot of tweets were talking about how to protect ourselves during this time this year.

Another source of negativity in this category is people’s feelings and experiences towards travel during this time, as can be indicated by the keyword ‘departure’. Many tweets were about how the virus has negatively impacted people’s travel experiences, commenting on all of the extra precautions in place to protect travelers from contracting the virus abroad and bringing it home. Lastly, a notable keyword is ‘results’ which refers to people getting tested for COVID-19. The fact that this word is so commonly spoken about in this category demonstrates how often it occurs, potentially hindering people from traveling or living their lives as usual, as well as well as potentially contributes to the large negative sentiment found in this group, where approximately 47% of the tweets hold negative sentiment and only around 5% of them are positive.

Government

There is evidence showing that in Canada, people have considered the government as playing an important role in supporting society since the start of the COVID-19 pandemic. In particular, research has shown that 52.2% were satisfied with the government’s public health response to the virus and 54.2% were satisfied with the governments’ economic response.[4] However, the results found in Figure 1,

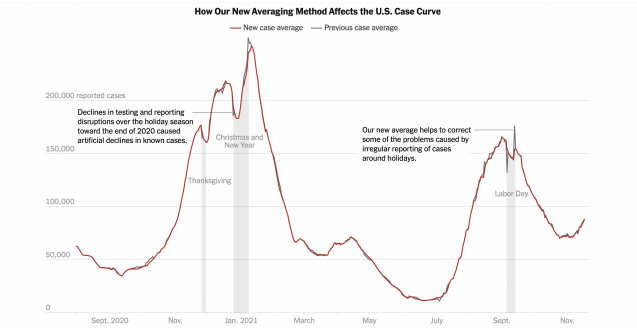

Figure 3: Evolution of COVID-19 cases in U.S. [6]

which contains the public's attitudes towards their respective government internationally, shows that in general people have been dissatisfied with their government's response to COVID-19. Specifically, this topic has the greatest proportion of negative sentiment out of all of the topics.

Since we have collected any COVID-related tweets in English, many of them are presumably from Americans, as we found that Donald Trump was been mentioned a lot, as well as the Republican party as seen through the keywords of 'donald', 'presidency' and 'gop'. We noticed that among 30 tweets that mentioned the former president of the United States, 29 of them are negative. This shows that the public has viewed Donald Trump's approach towards COVID-19 negatively.

The keyword 'masked' found in this category is presumably referring to mask-mandates being enforced by the government, and people's opinions on them. Since the majority of tweets in this group are negative, we may infer that in general people are not in favor. As well, the keywords of 'catastrophe' and 'ignorance' are clear negative connotations that people in general have towards the way their governments are protecting their citizens from the impacts of the virus. In addition, the word 'voters' in the list shows that the performance of the government during COVID-19 would greatly impact voters' choice on their choice of government during the next term.

Other

Although this category is not necessarily informative of a certain subject since it is a 'catch-all' group, we can still observe the associated keywords to try to explain how people feel about them. A notable keyword in this group is 'rogan', presumably referring to Joe Rogan: an American public figure who according to our analysis is a prominent figure in COVID-19 discourse. Other notable keywords in this category are 'war', 'lied', and 'caught' which seem negative in sentiment. However, we note that the majority of tweets in this group are neutral in sentiment, which may refer to the tweets in this category being more factual statements rather than being opinion-based.

Limitations of this Study

Even though we have made extensive conclusions based our analysis, we found that there are some limitations of the topics we designed which needs further investigation in order to make a more accurate conclusion. Firstly, the 'other' category includes around 40 tweets and this category is not informative. Thus, in future iterations we could include more topics instead of using 'other' category to uncover more useful information. Secondly, we used single annotation in this analysis. However, many tweets we have collected have mentioned more than one topic or had multiple sentiments. Conducting double annotation would generate more complete and informative results. As well, since our analysis is based on only 1000 tweets, it would be interesting to extend this experiment to a larger sample of tweets to get a more accurate depiction of how the general public feels about these topics.

Conclusion

Overall, through this exercise we have concluded that with respect to COVID-19 discourse through social media, in particular Twitter, people have been mainly talking about the themes of the virus and its variants and vaccination, as well as its effects on their daily lives, the government's role in limiting the spread, and the virus' societal implications, although to a lesser extent. In particular, people have talked about vaccines the most and it is the topic that possesses the highest proportion of positive sentiment. However, the new variant, Omicron, has been discussed a lot, especially how it may reduce the effectiveness of the vaccine, and that is why the vaccine booster was also a hot topic. In addition, the analysis shows that people resist vaccination mainly due to its potential risks involved. People have also been dissatisfied with their governments' approach and it is the topic that possesses the highest negative sentiment rate. Among all the tweets in this topic group, nearly half of them are negative and only 11% of the tweets are positive. In terms of its societal impact, keywords such as 'misinformation' and 'fight' stand out as strong themes, demonstrating the large negative association people have with how society has been affected by the virus. The virus has also negatively impacted people's daily lives through its effect on travel as well as celebrations. As a whole, the general sentiment is negative not only over the distribution of topics, but for each topic individually as well. We question whether this is due to the nature of the information collected, i.e. do people typically go on Twitter to express their dissatisfaction or frustrations? For future work into this topic, it would be interesting to see whether the discourse changes on other social media platforms such as Facebook or Reddit, as well as finding a way to limit the scope to only Canadian posts to see whether our findings are indeed accurate for Canadian society specifically.

Group Member Contributions

Sierra and Xuanzi were responsible for collecting the tweets and for producing data visualizations. Xuanzi determined the topics through open coding. All three of us took part in

annotating the tweets. Jaylene was responsible for computing tf-idf scores. Finally, all three of us contributed evenly to the report.

References

[1] Canada Health Infoway. 2021. Canadians' Health Care Experiences During COVID-19. https://www.infoway-inforoute.ca/en/component/edocman/3828-canadians-health-care-experiences-during-covid-19/view-document?Itemid=0.

[2] Government of Canada. 2021. Reported side effects following COVID-19 vaccination in Canada. https://health-infobase.canada.ca/covid-19/vaccine-safety/summary.html.

[3] Lidiane Cunha. 2021. Addressing the Impact of the COVID-19 Crisis on Our Healthcare System. https://telfer.uottawa.ca/en/research/innovative-thinking-rss/covid-19-healthcare-system/.

[4] McMaster University. 2021. COVID and Attitudes Towards the Role of Government. https://labourstudies.mcmaster.ca/research/impact-of-covid-19/attitudes-towards-government-covid.

[5] Parack, S. 2021. A comprehensive guide for using the Twitter API V2 with Tweepy in python. https://dev.to/twitterdev/a-comprehensive-guide-for-using-the-twitter-api-v2-using-tweepy-in-python-15d9.

[6] The New York Times. 2021. The Trouble With the Case Curve During the Holidays. https://www.nytimes.com/interactive/2021/11/22/us/covid-data-holiday-averages.html?auth=login-google1tap&login=google1tap.

[7] World Health Organisation. 2021. WHO Coronavirus (COVID-19) Dashboard. https://covid19.who.int.